DOI:10.11931/guihaia.gxzw202210078

白芷全基因组测序分析及 *BGLU* 基因家族分析

王雅兰，周罗静，张灵迂，章景，卞金辉，高继海*

（成都中医药大学西南特色中药资源国家重点实验室，成都 611137）

摘 要： 白芷为常用的药食同源的品种，既是临床常用中药，也是香料，用途十分广泛。为获取白芷全基因组序列信息，该研究用杭白芷叶片 DNA 为材料，采用 Nanopore 测序技术构建杭白芷全基因组数据库，并利用生物信息学方法对获得的核苷酸序列进行组装、功能注释以及进化分析研究。结果表明：（1）原始测序数据过滤后获得 662 Gb 三代数据，Read N50 约为 32 932 bp，经过组装得到杭白芷基因组大小为 5.6 Gb，Contig N50 约为 806 638 bp。（2）组装后的序列通过与 KOG、GO、KEGG 等功能数据库比对，得到了功能注释的基因共有 66.47%，KOG 功能注释结果表明杭白芷的蛋白功能主要集中在一般功能预测、翻译后修饰、蛋白质转换、伴侣以及信号转导机制；GO 功能分类表明杭白芷的基因集中在生物学过程及细胞组分；KEGG 通路注释表明参与代谢途径的基因占主要地位。（3）杭白芷的基因集中在 45 个 *BGLU* 家族基因。该研究首次利用第三代测序技术对杭白芷全基因组进行解析，为杭白芷的系统生物学研究奠定基础，有利于进一步深入开发和利用杭白芷，同时也对杭白芷中 *BGLU* 家族基因进行初步分析，为后续进一步研究 *BGLU* 在杭白芷生长发育中的功能提供了重要的理论基础。

关键词： 杭白芷，基因组，第三代测序技术，*BGLU* 基因家族，药用植物

中图分类号： Q943.2　　　　**文献标识码：** A

Complete genome sequencing and *BGLU* gene family analysis of *Angelica dahurica*

WANG Yalan, ZHOU Luojing, ZHANG Lingyu, ZHANG Jing, BIAN Jinhui, GAO Jihai*

(*Key Laboratory of Distinctive Chinese Medicine Resources in Southwest China*, *Chengdu University of Traditional Chinese Medicine*, Chengdu 611137, China)

Abstract: *Angelica dahurica* is a common kind of medicine and food homology, which is not only a common clinical traditional Chinese medicine, but also a spice, with a wide range of uses. In order to obtain the whole genome sequence information of *A. dahurica*, this study used *A. dahurica* var. *formosana* leaf DNA as material, and the Nanopore sequencing technology was used to establish its nucleotide sequences database, then genome assembly, function annotation and evolution analysis were carried out by bioinformatic methods. The results were as follows: (1) 662 Gb the third-generation data were obtained after the original sequencing data, with the Read N50

基金项目： 中央本级重大增减支项目（2060302）；国家中医药管理局项目（ZYYCXTD-D-202209）；四川省科技厅科技计划项目（2020YFN0152，22CXTD0009）；四川省中医药管理局项目（2022C001）；成都中医药大学人才提升项目（QNXZ2018017，QNXZ2019001）[Supported by Major Increase and Decrease of Expenditures of the Central Government (2060302); National Administration of Traditional Chinese Medicine (ZYYCXTD-D-202209); Science and Technology Program of Science and Technology Department of Sichuan Province (2020YFN0152, 22CXTD0009); Sichuan Provincial Administration of Traditional Chinese Medicine (2022C001); Talent Promotion of Chengdu University of Traditional Chinese Medicine Project (QNXZ2018017, QNXZ2019001)]。

第一作者： 王雅兰(1998-)，硕士研究生，研究方向为中药有效成分分析应用研究，（E-mail）wangyalan@stu.cdutcm.edu.cn。

***通信作者：** 高继海，博士，副教授，主要从事分子生药学研究，（E-mail）gaojihai@cdutcm.edu.cn。

about 32 932 bp. The assembled *A. dahurica* var. *formosana* genome size was 5.6 Gb, contig N50 being about 806 638 bp. (2) 66.47% of the genes from the assembled sequence got gene annotation after being compared with functional databases such as NR, KOG and KEGG. The result of KOG gene annotation was that the protein function of *A. dahurica* var. *formosana* concentrated in the general functional prediction only, posttranslational modification, protein turnover, chaperones signal transduction mechanisms. GO functional classification indicated that the genes of *A. dahurica* var. *formosana* concentrated on cell biological processes and components. KEGG analysis found that the *A. dahurica* var. *formosana* genes mostly involved in metabolic pathways. (3) 45 genes of *BGLU* family were identified in *A. dahurica* var. *formosana.* In this study, the whole genome of *A. dahurica* var. *formosana* is resolved by the third-generation sequencing technology for the first time, which lay a foundation for the systematic biological study of *A. dahurica* var. *formosana* , and is conducive to the further development and utilization. At the same time, the *BGLU* family genes were preliminarily analyzed, it also provides an important theoretical basis for the further study of the function of *BGLU* in the growth and development of *A. dahurica* var. *formosana* .

Key words: *Angelica dahurica* var. *formosana*, genome, the third-generation sequencing technology, *BGLU* gene family, medicinal plant

白芷为伞形科（Apiaceae）植物白芷（*Angelica dahurica*）或杭白芷（*A. dahurica* var. *formosana*）的干燥根，主产于四川、杭州等地，多为栽培品。白芷是常见的药食同源药材，在临床上可用于感冒头痛、眉棱骨痛、牙痛、疮疡肿痛等各种类型的疼痛症状(国家药典委员会，2020)，在日常生活中也可以作为香料使用。同时，由于其气味芳香，还被广泛应用于化妆品、洗护用品等方面(于静等，2014)。白芷含有多种活性成分，如香豆素类、挥发油类、多糖类、生物碱类等(Li et al., 2014; Zhao et al., 2022)，现代研究表明其主要有效成分是香豆素类和挥发油类，具有解热镇痛、抗炎、抗病原微生物、抗肿瘤、降压、保肝等多种药理作用(吉庆等，2020；王蕊等，2020)。

白芷的应用前景十分广泛，但近年对白芷的研究多数集中在化学成分、栽培技术、药理药效的解析等方面，而少有关于白芷遗传信息的研究，目前只见对白芷转录组进行测序分析(吴萍等，2020)的研究，对白芷 COSNTANS-like(蒋翼杰等，2021)、NAC(黄文娟等，2021)、MYB-related(姚菲等，2022)基因家族的研究以及白芷中香豆素合成关键基因的挖掘均是依据转录组数据进行(刘洋，2019)，白芷基因组数据的缺乏导致无法获取白芷完整的遗传信息，更多的研究无法开展或进一步深入，故对其进行全基因组测序对是十分重要的。

香豆素类成分既是白芷的药效成分又是香气成分，香豆素类化合物广泛存在于自然界的多种植物中，如伞形科、芸香科、桑科等的植物(Venugopala et al., 2013)，近年来对香豆素的生物合成途径研究较多，一些关键酶及功能作用的解析也较为清晰(段珍等，2022)。其中就包括β-葡萄糖苷酶（β-glucosidase, *BGLU*），β-葡萄糖苷酶家族不仅在香豆素的生物合成中起到重要调控作用，而且广泛参与植物激素信号激活(Sun et al., 2014)、次生代谢(Sampedro et al., 2017)等多种重要生理过程。有研究表明β-葡萄糖苷酶家族在草木樨的香豆素合成中起到重要调控作用(吴凡，2021)，在玉米中能通过催化碳水化合物部分和香豆素核心结构间的β-葡萄糖苷键的水解，进而产生香豆素苷元形式；黑曲霉来源的β-葡萄糖苷酶对丁公藤粗提物中的东莨菪苷可特异性水解，并使其含量提高 47%(于坤朋等，2023)；从拟南芥中分离的 3 种β-葡萄糖苷酶能特异性水解东莨菪苷。东莨菪苷在β-葡萄糖苷酶的作用下水解成东莨菪内酯，东莨菪内酯属于香豆素类成分，在白芷中也有存在，课题组推

测在白芷的香豆素成分合成中，*BGLU* 基因也起到关键作用。

由于目前未见关于白芷高质量基因组的研究，对白芷中香豆素合成途径的解析也较少，为了进一步丰富白芷的遗传进化的研究资料，本研究通过对杭白芷进行第二代、三代基因组测序，对测序数据进行组装、注释等，获得杭白芷的高质量基因组，并进行功能注释、基因家族聚类等分析，然后挖掘香豆素合成途径关键基因 *BGLU*，通过在线软件对基因组中提取的 *BGLU* 序列进行基本的特征分析，拟探讨以下问题：(1)杭白芷基因组概况；(2)基因功能主要集中在哪些生物学过程及代谢通路；(3)*BGLU* 基因家族的基本特征是什么。本研究将为白芷的后续研究提供数据基础及分子基础，能为后续深入研究 *BGLU* 基因家族在白芷香豆素合成途径中的功能提供前期基础。

1 材料与方法

1.1 材料及 DNA 提取

杭白芷植株来自成都中医药大学药用植物园，经国家中药种质资源库专家高继海副教授鉴定为伞形科植物杭白芷（*Angelica dahurica* var. *formosana*）。采摘新鲜、幼嫩、无病虫害的叶片，先用蒸馏水清洗表面，然后使用 75%乙醇清洗 3 次，擦干后置于-80 ℃冻存，备用。

参照沙丽萍(2018)采用 CTAB 法提取杭白芷叶片 DNA。提取的 DNA 需通过琼脂糖凝胶电泳和 Qubit Fluorometer 检测浓度，以及 Nanodrop 检测纯度和完整度。

1.2 文库构建及测序

（1）MGISEQ-200 测序：提取的杭白芷基因组 DNA 经检测合格以后，通过酶解随机打断成片段，经末端修复、加 A 尾、加测序接头、纯化、PCR 扩增等步骤构建插入片段长度为 150 bp 的 DNA 文库。将构建好的文库在 MGISEQ-200 平台进行双端测序。

（2）Nanopore 测序：利用磁珠对检测合格的 DNA 进行富集和纯化，对纯化后的 DNA 进行损伤修复、末端修复、加 A 尾后再纯化，将产物进行测序相关的连接及纯化，得到最终上机文库，用 Qubit 对建好的 DNA 文库进行精确的定量检测，取一定量的 DNA 文库混合上机相关试剂后加入流动槽中，在 GridION 测序仪上进行单分子测序，得到原始数据。

1.3 基因组测序数据的质量控制

二代原始测序数据中包含的接头信息，低质量碱基，未测出的碱基（以 N 表示）等会对后续的信息分析造成很大的干扰，这些干扰信息需要利用 FastQC v0.11.9 软件和 Trimmomatic v0.39 软件进行过滤，最终得到有效数据（clean reads）用于后续分析。

对于三代 Nanopore 测序数据使用 NanoPlot v1.20.0 软件对测序质量进行检测，再利用 NanoFlit v2.8.0 软件进行低质量和短片段数据的过滤。

1.4 基因组大小和杂合度评估

利用 MGISEQ-200 测序得到的 reads 数据，采用 Jellyfish v1.1.10 做 Survey 分析来预估基因组大小，杂合率、及重复序列占比，以判断基因组复杂情况。采用基因 K-mer 的分析方法来估计杭白芷基因组特征。

1.5 基因组组装及评估

为得到高准确性的三代组装结果，先采用 Canu v2.1.1(Koren et al., 2017)软件对 Clean Data 进行纠错，然后将纠错后的数据进行组装，用 Racon v1.0.0(Senol et al., 2019)软件对组装结果进行纠错，再用 Pilon v1.22 软件使用二代数据进行校正，最后利用 BUSCO v5.1.2(Simão et al., 2015)软件对组装完成的基因组进行完整性评估。

1.6 序列预测

首先，基于结构预测和从头预测(*Ab initio*)的原理，使用 LTR Finder v1.05(Xu et al.,

2007)、RepeatScout v1.0.6、PILER-DF v2.4 软件构建重复序列数据库，利用 PASTEClassifier v2.0 对构建好的重复序列库进行分类；然后，基于重复序列数据库 Repbase(https://www.girinst.org/repbase/)合并作为最终的杭白芷基因组的重复序列数据库；最后，基于构建好的数据库采用 RepeatMasker v4.1.2 软件对杭白芷进行重复序列的预测。

基于从头预测和同源物种预测(Homolog)两种原理对杭白芷基因组进行基因预测，并对预测结果进行评估。首先，利用 Genscan v1.0、Augustus v3.3.1、GlimmerHMM v3.0.4、GeneID v1.4、SNAP v8.0.0 进行从头预测；然后，使用 GeMoMa v1.3.1 进行基于同源物种的预测；最后，利用 EvidenceModeler v1.1.0 整合、校正上述方法得到的预测结果。针对非编码 RNA 预测，包括了 microRNA、rRNA 及 tRNA 等已知功能的 RNA，分别基于 Rfam(Finn et al., 2006)数据库和 miRBase 数据库并利用 Infenal v1.1.3 进行 rRNA 和 microRNA 预测；利用 tRNAscan-SE v2.0.7 识别 tRNA。

1.7 功能基因注释

对预测得到的基因序列与 NR(Non-Redundant Protein Database)、KOG(EuKaryotic Orthologous Groups)、KEGG(Kyoto Encyclopedia of Genes and Genomes)、TrEMBL 等功能数据库做 BLAST v2.2.31 比对，设置比对筛选阈值(e-value＜1e-5)，得到基因功能注释。基于 NR 数据库比对结果，应用软件 Blast2GO v5.2.5 进行 GO 数据库的功能注释。

1.8 基因家族聚类分析及系统进化分析

利用杭白芷和其同科物种的对比来寻找基因家族，首先从 NCBI 数据库中下载杭白芷同科植物芹菜(*Apium graveolens*)(Song et al., 2021)、胡萝卜(*Daucus carota* subsp. *sativus*)(Iorizzo et al., 2016)的蛋白序列，从 CGDB(http://cgdb.bio2db.com)下载芫荽(*Coriandrum sativum*)(Song et al., 2020)蛋白序列。通过 OrthoMCL v2.0(Li et al., 2003)软件对 all-vs-all blastp 获得的所有物种蛋白序列间的相似性关系进行聚类分析。通过从 OrthoMCL 聚类结果中提取的单拷贝蛋白序列，再 Muscle v3.8.31(Edgar, 2004)软件中进行对比，再通过 RAxML v8.2.12(Guindon & Gascuel, 2003)软件采用最大似然法(ML TREE)构建进化树。

1.9 杭白芷 *BGLU* 基因家族成员挖掘

利用 SMART 数据库，获得拟南芥 *BGLU* 基因家族的典型结构域序列 tBLASTN (*P*=0.001)，并搜索杭白芷基因组数据库，通过 Pfam 数据库得到杭白芷中所有 *BGLU* 基因家族成员。

1.10 *BGLU* 家族基因理化性质、亚细胞定位、蛋白二级结构及保守域分析

利用 ProtParam tool(https://web.expasy.org/protparam/)(Wilkins et al., 1999)在线软件对 BGLU 家族蛋白其进行理化性质分析；用 Plant-mPLoc(http://www.csbio.sjtu.edu.cn/bioinf/plant-multi/) 及 WoLF PSORT(https://wolfpsort.hgc.jp/)在线软件综合分析其亚细胞定位；使用 SOMPA(https://npsa-prabi.ibcp.fr/cgi-bin/npsa_automat.pl?page=npsa_sopma.html)在线软件分析其二级结构；通过 MEME(https://meme-suite.org/meme/tools/meme)在线软件分析保守结构域。

1.11 *BGLU* 家族系统进化分析

利用 MEGA 软件中的 Clustal W v2.0(Larkin et al., 2007)程序对杭白芷和拟南芥的 BGLU 家族蛋白序列进行对比，将对比结果采用邻接法构建系统发育树。

2 结果与分析

2.1 基因组测序

通过测序平台对杭白芷叶片进行全基因组测序，对原始数据的 reads 质量值进行初步过滤，去掉低质量和短片段的 reads，统计得到 150 Gb 二代原始数据，得到 662 Gb 三代原始数据。三代数据中，Read N50 为 32 932 bp，最长 reads 的长度为 422 833 bp，平均长度

为 27 750 bp，测序质量符合后续组装要求。Survey 分析得出杭白芷基因组的大小约为 5.2 Gb。

2.2 基因组组装及评估

借助 Canu 软件对杭白芷进行纠错组装，基因组大小约为 5.6 Gb，Contig N50 为 806 638 bp，最长的 Contig 为 21 677 961 bp，GC 含量为 35.73%。组装后的基因组采用 BUSCO v5.1.2 软件评估，在组装的基因中共找到 1 580 个完整的 BUSCO 基因，其中完整单拷贝的 1 272 个，Fragmented BUSCO 18 个基因，有 16 个基因在 Embryophyta_odb10 数据库中没有找到，BUSCO 评估基因组完整度为 97.9%，表明该组装结果较为完整。

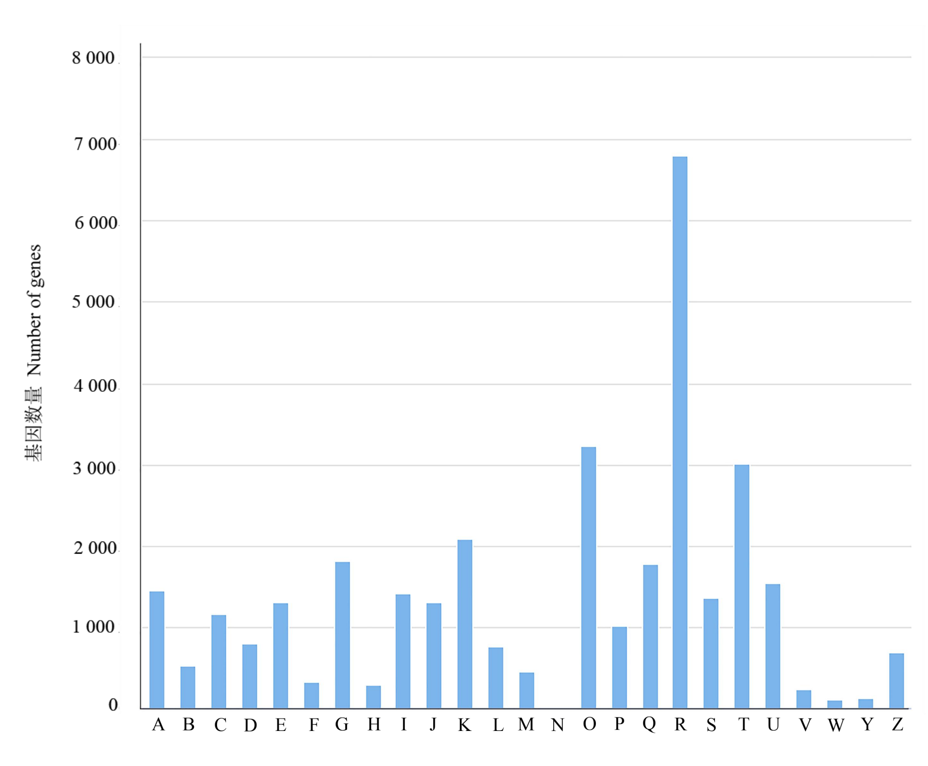
2.3 基因预测结果

利用 RepeatMasker v4.1.2 软件进行重复序列预测得到包含 5.4 Gb 重复序列的杭白芷基因组，占比 91.36% 。其中长散在重复序列(LINE)数目为 21 726 条，占比 0.41%；短散在重复序列(SINE)数目为 0 条；长末端重复序列(LTR)数目为 3 550 524 条，占比 69.07 %，copyia 数目为 1 083 004 条，占比 30.01%，gypsy 数目为 989 985 条，占比 24.56%，Roling-circles 数目为 2 893 条，占比 0.03%；简单重复序列(SSR)数目为 7 710 条，占比 0.03%。

在获得的 67 004 个基因中，有 34 119 (93.1%)个基因得到了其他物种同源性鉴定或 RNA-seq 数据的支持。共鉴定出 2 749 个非编码 RNA (ncRNA)，包括 20 个核糖体 RNA (rRNA)、781 个转移 RNA (tRNA)、97 个小分子 RNA (microRNA)和 15 505 个小核 RNA (snRNA)。

2.4 基因功能注释与分析

通过 KOG 功能注释(图 1)可得出，杭白芷基因组共 29 788 个基因获得注释，占预测到的总基因数的 44.46%。从图中可以看出，杭白芷的蛋白功能主要集中在 次级代谢产物的生物合成，转运和代谢，占比为 10.8%，其次为信号转导机制，占比为 10.1%，转录，占比为 6.7%，碳水化合物转运和代谢，占比为 3.7%；一般功能预测，占比为 22.8%。这些基因的差异性表达可以为今后杭白芷的深入研究提供数据支持。



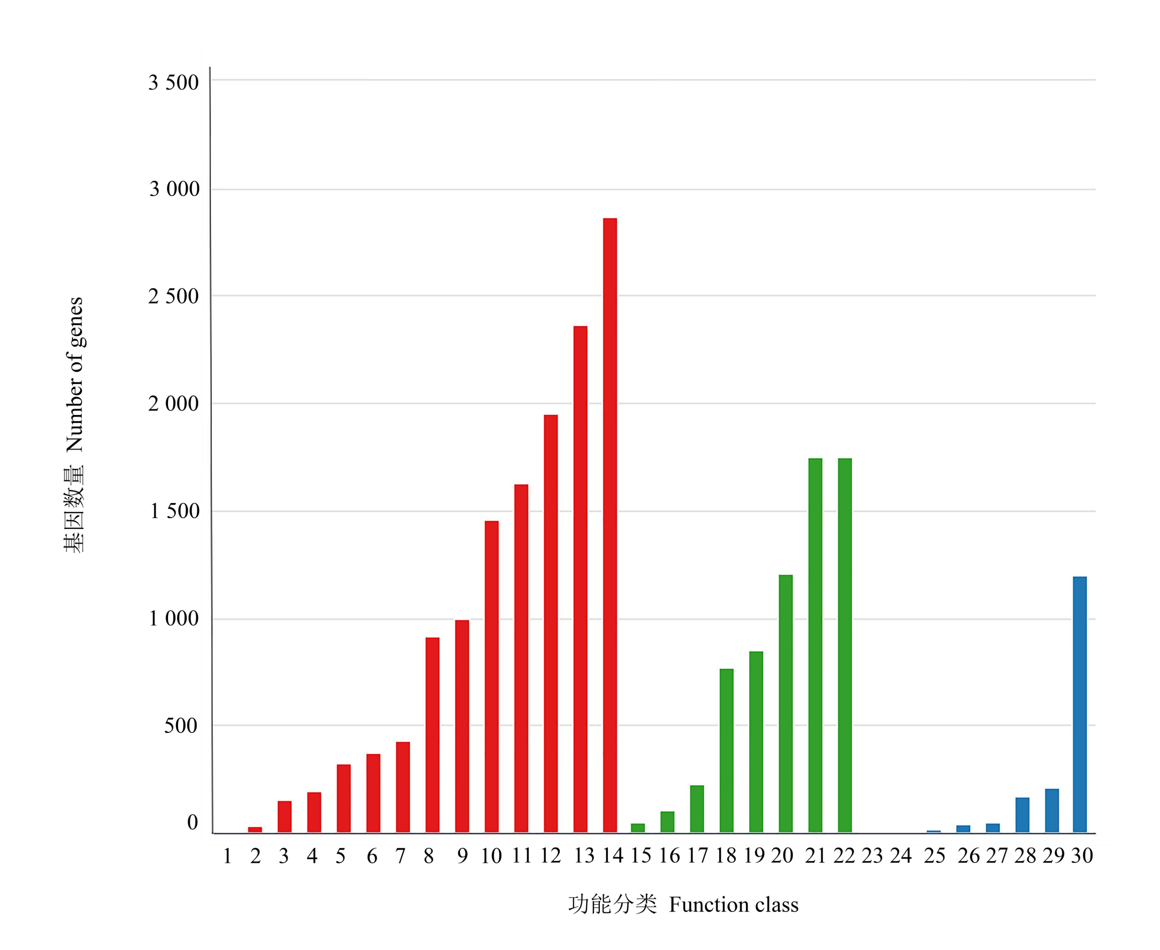
A. RNA 加工和修饰；**B**. 染色质结构和动力学；**C**. 能量生产和转换；**D**. 细胞周期调控，细胞分裂，染色体分配；**E**. 氨基酸转运和代谢；**F**. 核苷酸转运和代谢；**G**. 碳水化合物转运和代谢；**H**. 辅酶转运和代谢；**I**. 脂质转运和代谢；**J**. 翻译，核糖体结构和生物合成；**K**. 转录；**L**. 复制，重组和修复；**M**. 细胞壁/细胞膜/胞外被膜生物合成；**N**. 细胞运动；**O**. 翻译后修饰，蛋白质转换，伴侣；**P**. 无机离子转运和代谢；**Q**. 次级代谢产物的生物合成，转运和代谢；**R**. 一般功能预测；**S**. 功能未知；**T**. 信号转导机制；**U**. 胞内运输，分泌和囊泡运输；**V**. 防御机制；**W**. 胞外结构；**Y**. 细胞核结构；**Z**. 细胞骨架。

A. RNA processing and modification; **B**. Chromatin structure and dynamics; **C.** Energy production and conversion; **D**. Cell cycle control, cell division, chromosome partitioning; **E**. Amino acid transport and metabolism; **F**. Nucleotide transport and metabolism; **G**. Carbohydrate transport and metabolism; **H**. Coenzyme transport and metabolism; **I**. Lipid transport and metabolism; **J**. Translation, ribosomal structure and biogenesis; **K**. Transcription; **L**. Replication, recombination and repair; **M**. Cell wall/membrane/envelope biogenesis; **N**. Cell motility; **O**. Posttranslational modification, protein turnover, chaperones; **P**. Inorganic ion transport and metabolism; **Q**. Secondary metabolites biosynthesis, transport and catabolism; **R**. General function prediction; **S**. Function unknown; **T**. Signal transduction mechanisms; **U**. Intracellular trafficking, secretion, and vesicular transport; **V**. Defense mechanisms; **W**. Extracellular structures; **Y**. Nuclear structure; **Z**. Cytoskeleton.

图 1 KOG 功能分类注释图

Fig.1 KOG function annotation classification chart

杭白芷基因组 GO 注释(图 2)表明，共有 44 540 个基因具有 GO 注释功能，占预测到的总基因数的 66.47%。功能主要分布在生殖、细胞过程、胁迫应答、细胞、细胞部位等的基因占优势，其中在生殖的基因占比最多。



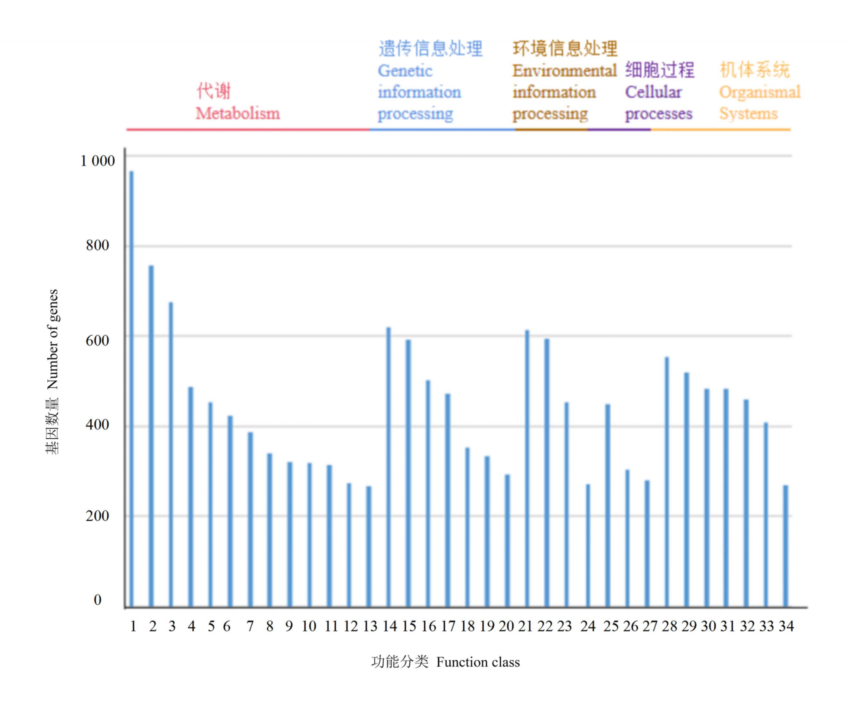
1. 细胞杀伤；**2**. 多生物过程；**3**. 多细胞生物过程；**4**. 信号转导；**5**. 免疫系统过程；**6**. 定位；**7**. 发育过程；**8**. 细胞组分或生物合成；**9**. 生物调控；**10**. 单一生物过程；**11**. 代谢过程；**12**. 胁迫应答；**13**. 细胞过程；**14**. 生殖；**15**. 细胞膜部位；**16**. 膜封闭腔；**17**. 细胞膜；**18**. 细胞器部位；**19**. 高分子复合物；**20**. 细胞器；**21**. 细胞；**22**. 细胞部位；**23**. 分子传感器活动；**24**. 分子功能调节器；**25**. 转录因子活性，蛋白结合；**26**. 转运活性；**27**. 信号转导活性；**28**. 核酸结合转录因子活性；**29**. 催化活性；**30**. 结合。
图中红色框代表生物学过程，绿色框代表细胞组分，蓝色框代表分子功能。

1. Cell killing; **2**. Multi-organism process; **3**. Multicellular organismal process; **4**. Signaling; **5**. Immune system process; **6**. Localization; **7**. Developmental process; **8**. Cellular component organization or biogenesis; **9**. Biological regulation; **10**. Single-organism process; **11**. Metabolic process; **12**. Response to stimulus; **13**. Cellular process; **14**. Reproduction; **15**. Membrane part; **16**. Membrane-enclosed lumen; **17**. Membrane; **18**. Organelle part; **19**. Macromolecular complex; **20**. Organelle; **21**. Cell; **22.** Cell part; **23**. Molecular transducer activity; **24**. Molecular function regulator; **25**. Transcription factor activity, protein binding; **26**. Transporter activity; **27**. Signal transducer activity; **28**. Nucleic acid binding transcription factor activity; **29**. Catalytic activity; **30**. Binding.
The red represents the biological process, the green represents the cellular components, and the blue represents the molecular function in this figure.

图 2 GO 注释分类图

Fig.2 GO annotation classification chart

KEGG 通路注释(图 3)对杭白芷的 15 263 个基因进行了通路注释，占预测到的总基因数的 22.78%。其注释结果表明参与“代谢”的基因占主要，其中微生物在不同环境中的代谢、碳代谢、氨基酸生物合成为主要代谢通路。



1. 微生物在不同环境中的代谢；**2**. 碳代谢；**3**. 氨基酸的生物合成；**4**. 苯丙素的生物合成；**5**. 嘌呤代谢；**6**. 淀粉和蔗糖代谢；**7**. 糖酵解/糖异生；**8**. 氧化磷酸化；**9**. 嘧啶代谢；**10**. 戊糖、葡萄糖醛酸转换；**11**. 氨基糖和核苷酸糖代谢；**12**. 甘油磷脂代谢；**13**. 色氨酸代谢；**14**. 内质网中的蛋白质加工；**15**. 剪接体；**16**. 核糖体；**17**. 核质运输；**18**. 泛素介导的蛋白水解；**19**. mRNA 监测通路；**20**. RNA 降解；**21**. MAPK 信号通路；**22**. 植物激素信号转导；**23**. NF-kappa B 信号通路；**24**. PI3K-Akt 信号通路；**25**. 胞吞；**26**. 细胞周期；

27. 卵母细胞减数分裂；**28**. 生成信号通路；**29**. NOD 样受体信号通路；**30**. 植物-病原互作；**31**. Toll 样受体信号通路；**32**. Toll 和 Imd 信号通路；**33**. 生热作用；**34**. 胰岛素信号通路。

1. Microbial metabolism in diverse environments; **2**. Carbon metabolism; **3**. Biosynthesis of amino acids; **4**. Phenylpropanoid biosynthesis; **5**. Purine metabolism; **6**. Starch and sucrose metabolism; **7**. Glycolysis / Gluconeogenesis; **8**. Oxidative phosphorylation; **9**. Pyrimidine metabolism; **10**. Pentose and glucuronate interconversions; **11**. Amino sugar and nucleotide sugar metabolism; **12**. Glycerophospholipid metabolism; **13**. Tryptophan metabolism; **14**. Protein processing in endoplasmic reticulum; **15**. Spliceosome; **16**. Ribosome; **17**. Nucleocytoplasmic transport; **18**. Ubiquitin mediated proteolysis; **19**. mRNA surveillance pathway; **20**. RNA degradation; **21**. MAPK signaling pathway; **22**. Plant hormone signal transduction; **23**. NF-kappa B signaling pathway; **24**. PI3K-Akt signaling pathway; **25**. Endocytosis; **26**. Cell cycle; **27**. Oocyte meiosis; **28**. Neurotrophin signaling pathway; **29**. NOD-like receptor signaling pathway; **30**. Plant-pathogen interaction; **31**. Toll-like receptor signaling pathway; **32**. Toll and Imd signaling pathway; **33**. Thermogenesis; **34**. Insulin signaling pathway.

图 3 KEGG 功能注释图

Fig.3 KEGG function annotation diagram

2.5 基因家族聚类分析及系统进化分析

将杭白芷与同科植物芫荽、芹菜、胡萝卜的蛋白序列进行对比，在杭白芷基因组的 67 004 个蛋白序列中共鉴定出 23 151 个基因家族，其中 4 004 个基因家族包含 18 151 个基因特异存在于杭白芷中，4 种植物所共有的基因家族有 1 030 个（图 4）。

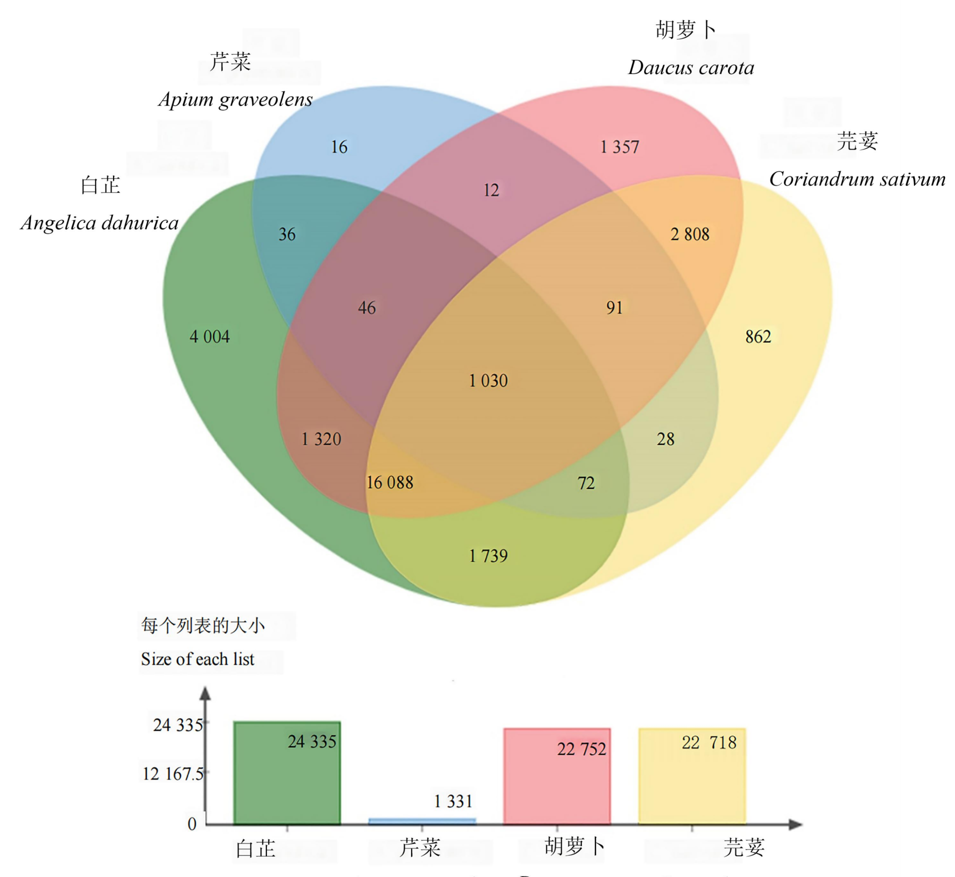


图 4 4 个基因家族 Venn 图

Fig.4 Venn diagram of gene families of 4 species

为进一步研究杭白芷的种属关系，以 96 条单拷贝蛋白序列进行比较分析，选择拟南芥(*Arabidopsis thaliana*)、玉米(*Zea mays*)、无油樟(*Amborella trichopoda*)以及同为伞形科的芫荽、芹菜、胡萝卜、当归(*Angelica sinensis*)共 7 个已知基因组信息的物种，与杭白芷构建遗传进化树（图 5）。结果表明杭白芷与芫荽聚为一支，两物种间亲缘关系较近。

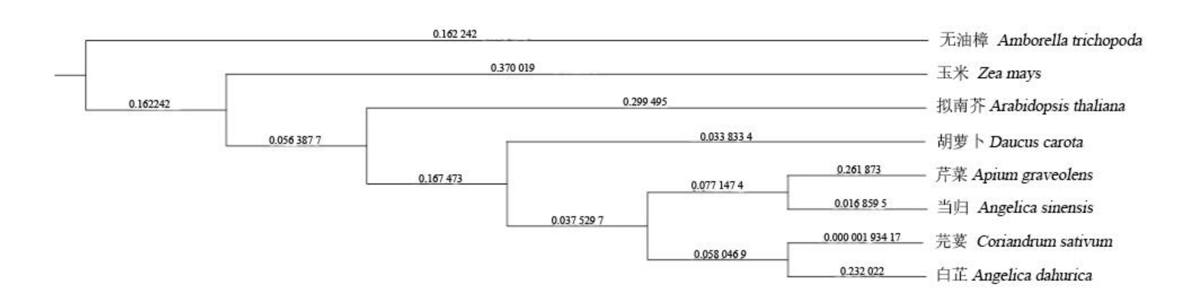


图 5 物种间进化关系

Fig.5 Evolutionary relationships between species

2.6 杭白芷 *BGLU* 家族基因理化性质、亚细胞定位分析

在杭白芷全基因组中共鉴定到 45 个 *BGLU* 家族基因，分别命名为 *AdBGLU*01~*AdBGLU*45，利用 Protparam Tool 进行理化性质分析，Plant-mPLoc 及 WoLF PSORT 进行亚细胞定位（表 1）。结果表明，杭白芷的 45 个 *BGLU* 基因编码的氨基酸数目在 51~930 之间，最长包含 930 个氨基酸残基（*AdBGLU*32），最短包含 51 个氨基酸残基（*AdBGLU*30）；不稳定指数在 11.18~61.86 之间，其中 38 个的不稳定系数小于 40，推测其为稳定蛋白，其余 7 个为不稳定蛋白；脂肪系数为 56.76~113.25，说明蛋白的热稳定性较好；亲水系数在-0.643~0.35 之间，其中 7 个为正值，38 个为负值，说明主要为亲水性蛋白；等电点在 4.24~10.35 之间，说明氨基酸大多为弱酸或弱碱性；亚细胞定位预测结果将 *AdBGLU* 家族成员分别定位于细胞核、细胞质、叶绿体、液泡中。*AdBGLU* 基因家族的不同成员间理化性质差异较大，且亚细胞定位较多，推测该基因家族成员功能较为多样，在生物体内参与不同的生理过程。

表 1 杭白芷 BGLU 家族蛋白理化性质、亚细胞定位

Table1 Physicochemical properties and subcellular localization of AdBGLU

编号 Number	基因 ID Gene ID	氨基酸数目 Number of amino acids	相对分子质量 Molecular weight	等电点 Theoretical pI	原子总数 Total number of atoms	不稳定指数 Instability index	脂肪系数 Aliphatic index	平均亲疏水性 Grand average of hydropa-thicity	亚细胞定位 Subcellular location
*AdBGLU*01	>evm.model.contig_30308_np12.2	86	9 761.19	10.35	1 388	40.46	91.74	-0.277	细胞核 Nucleus
*AdBGLU*02	>evm.model.contig_30308_np12.3	57	6 466.10	4.62	889	15.95	82.11	-0.261	细胞质 Cytoplasm

*AdBGLU*03	>evm.model.contig_28255_np12.3	641	72 592.77	6.02	10 144	37.66	80.66	-0.291	叶绿体 Chloroplast
*AdBGLU*04	>evm.model.contig_52431_np12.2	259	2 9687.46	8.44	4 171	35.24	96.25	0.054	叶绿体 Chloroplast
*AdBGLU*05	>evm.model.contig_52794_np12.2	512	58 122.58	5.89	8 067	28.22	77.93	-0.276	液泡 vacuole
*AdBGLU*06	>evm.model.contig_4149_np12.12	492	56 116.18	5.49	7 776	31.44	79.86	-0.381	液泡 vacuole
*AdBGLU*07	>evm.model.contig_3487_np12.4	641	72 634.86	6.10	10 153	38.25	80.66	-0.293	叶绿体 Chloroplast
*AdBGLU*08	>evm.model.contig_2826_np12.18	479	55 784.70	5.76	7 698	22.53	69.98	-0.506	细胞核 Nucleus
*AdBGLU*09	>evm.model.contig_6813_np12.9	531	60 685.64	5.28	8 387	36.57	77.10	-0.310	液泡 vacuole
*AdBGLU*10	>evm.model.contig_15195_np12.3	54	6 007.75	4.52	845	19.51	99.26	-0.019	细胞质 Cytoplasm
*AdBGLU*11	>evm.model.contig_5102_np12.18	58	6 671.41	4.62	920	18.04	85.69	-0.074	细胞质 Cytoplasm
*AdBGLU*12	>evm.model.contig_41190_np12.1	475	54 912.12	6.44	7 643	35.78	80.11	-0.419	叶绿体 Chloroplast
*AdBGLU*13	>evm.model.contig_51569_np12.1	249	28 330.9	5.04	3 959	28.72	81.41	-0.39	液泡 vacuole
*AdBGLU*14	>evm.model.scaffold_848_np12.18	253	28 731.25	7.09	4 012	44.65	89.72	-0.019	细胞质 Cytoplasm
*AdBGLU*15	>evm.model.contig_13620_np12.1	189	21 311.79	5.01	3 002	30.18	111.27	0.35	叶绿体 Chloroplast
*AdBGLU*16	>evm.model.contig_875_np12.20	310	35 326.65	7.09	4 982	34.39	94.32	-0.22	细胞质 Cytoplasm
*AdBGLU*17	>evm.model.contig_5591_np12.41	73	8 246.25	4.4	1 148	35.45	78.9	-0.156	细胞质 Cytoplasm
*AdBGLU*18	>evm.model.contig_5591_np12.42	80	9 061.28	5.04	1 275	23.91	99.87	-0.085	细胞质 Cytoplasm
*AdBGLU*19	>evm.model.contig_7151_np12.13	157	17 864.13	6.04	2 465	26.02	73.89	-0.346	细胞质 Cytoplasm
*AdBGLU*20	>evm.model.contig_4524_np12.5	62	7 085.97	5.8	985	26.27	81.94	-0.324	细胞核 Nucleus
*AdBGLU*21	>evm.model.contig_6554_np12.21	52	5 934.62	4.76	829	18.43	99.23	-0.448	细胞质 Cytoplasm
*AdBGLU*22	>evm.model.contig_8631_np12.4	473	52 735.38	6.57	7 307	23.91	72.98	-0.364	细胞质 Cytoplasm
*AdBGLU*23	>evm.model.contig_8631_np12.5	518	58 661.26	5.45	8 155	28.05	82.82	-0.237	液泡 vacuole
*AdBGLU*24	>evm.model.contig_26865_np12.1	132	14 503.44	9.84	2 034	52.37	68.64	-0.495	叶绿体 Chloroplast
*AdBGLU*25	>evm.model.contig_8414_np12.6	107	12 102.53	5.09	1 661	11.18	80.28	-0.342	液泡 vacuole
*AdBGLU*26	>evm.model.contig_9063_np12.8	281	31 082.92	4.84	4 352	34.54	91.28	0.025	细胞核 Nucleus

*AdBGLU*27	>evm.model.contig_16310_np12.17	638	73 116.7	7.08	10 241	32.27	83.71	-0.289	叶绿体 Chloroplast
*AdBGLU*28	>evm.model.contig_5761_np12.10	511	58 131.95	8.07	8 096	29.65	80.94	-0.264	液泡 vacuole
*AdBGLU*29	>evm.model.contig_4290_np12.7	74	8 391.65	7.00	1 140	52.63	56.76	-0.043	细胞质 Cytoplasm
*AdBGLU*30	>evm.model.contig_9965_np12.1	51	5 765.41	4.24	800	18.36	93.53	-0.271	叶绿体 Chloroplast
*AdBGLU*31	>evm.model.contig_5955_np12.1	396	44 314.95	5.76	6 198	30.14	78.81	-0.391	叶绿体 Chloroplast
*AdBGLU*32	>evm.model.contig_62681_np12.3	930	106 242.47	6.17	14 772	36.27	74.69	-0.486	细胞质 Cytoplasm
*AdBGLU*33	>evm.model.contig_3403_np12.17	121	13 148.77	4.78	1 827	38.14	80.5	0.043	细胞质 Cytoplasm
*AdBGLU*34	>evm.model.contig_15768_np12.11	507	58 332.13	6.81	8 127	37.22	81.79	-0.371	叶绿体 Chloroplast
*AdBGLU*35	>evm.model.contig_9908_np12.3	77	8 591.58	4.64	1 178	11.55	83.64	-0.196	细胞质 Cytoplasm
*AdBGLU*36	>evm.model.contig_20919_np12.2	295	33 829.6	6.75	4 802	26.76	113.25	0.259	细胞质 Cytoplasm
*AdBGLU*37	>evm.model.contig_20919_np12.4	321	36 622.81	6.75	5 194	25.13	110.44	0.214	叶绿体 Chloroplast
*AdBGLU*38	>evm.model.contig_3686_np12.4	510	59 022.16	8.89	8 224	22.41	73.59	-0.456	细胞质 Cytoplasm
*AdBGLU*39	>evm.model.contig_4890_np12.4	303	33 883.74	9.21	4 740	37.2	79.41	-0.257	细胞质 Cytoplasm
*AdBGLU*40	>evm.model.contig_10929_np12.6	228	25 724.04	6.42	3 638	29.16	106.75	0.226	叶绿体 Chloroplast
*AdBGLU*41	>evm.model.contig_40130_np12.2	85	9 541.12	9.74	1 368	43.04	111.18	-0.048	细胞质 Cytoplasm
*AdBGLU*42	>evm.model.contig_9053_np12.1	177	20 452.06	5.85	2 831	51.96	78.14	-0.401	叶绿体 Chloroplast
*AdBGLU*43	>evm.model.contig_4791_np12.1	78	8 918.13	5	1 242	26.56	93.85	-0.228	叶绿体 Chloroplast
*AdBGLU*44	>evm.model.contig_5084_np12.19	511	58 990.4	9.18	8 224	22.79	71.02	-0.442	叶绿体 Chloroplast
*AdBGLU*45	>evm.model.scaffold_6091_np12.6	269	31 269.15	9.95	4 431	61.86	78.25	-0.643	细胞核 Nucleus

2.7 杭白芷 BGLU 基因家族蛋白二级结构及保守域分析

在线分析网站对杭白芷 BGLU 家族蛋白的二级结构分析表明（表 2），BGLU 家族中α-螺旋和无规则卷曲所占比例最大，其中α-螺旋所占比例最大的有 27 个，无规则卷曲所占比例最大的有 18 个。无规则卷曲为蛋白中的不稳定编码区，因此可推测无规则卷曲越多，该家族成员的功能越多样(姚菲等，2022)。

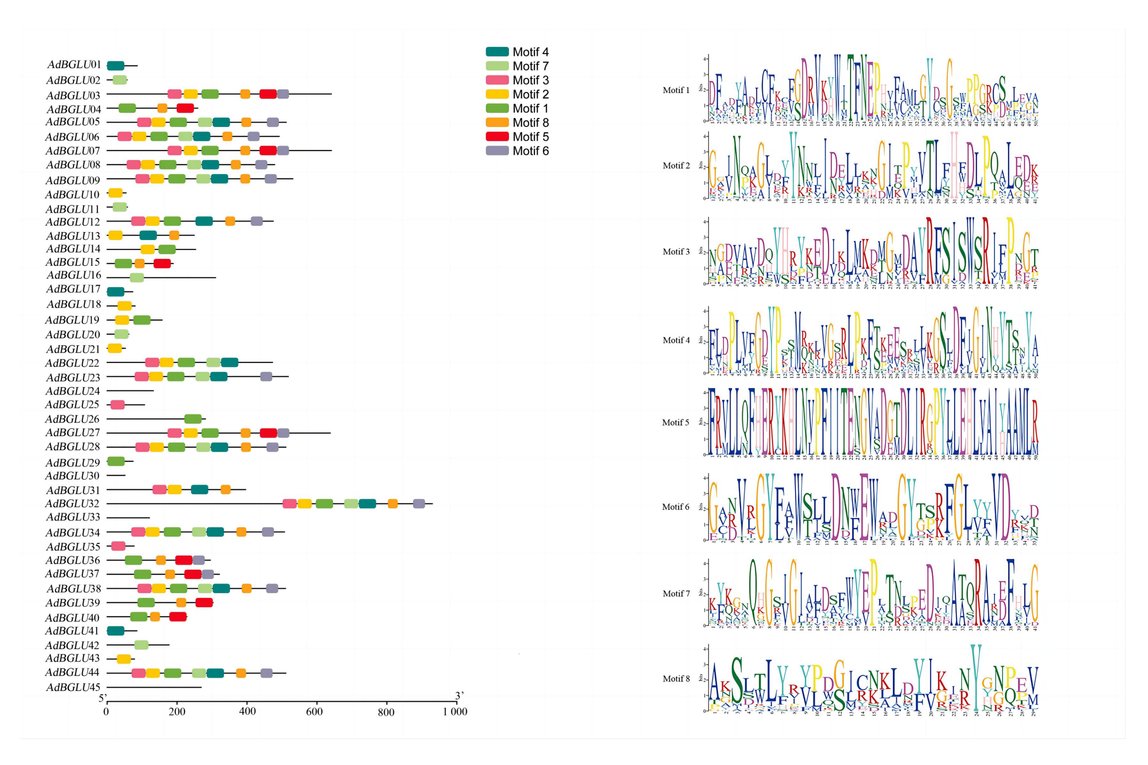
表 2 杭白芷 BGLU 家族蛋白二级结构分析

Table 2 Analysis of secondary structure of AdBGLU family proteins

编号 Number	基因 ID Gene ID	α-螺旋 Alpha	β-转角 Beta	延伸链 Extended	无规则卷曲

		helix	turn	strand	Random coil
*AdBGLU*01	>evm.model.contig_30308_np12.2	30.23	5.81	20.93	43.02
*AdBGLU*02	>evm.model.contig_30308_np12.3	28.07	3.51	33.33	35.09
*AdBGLU*03	>evm.model.contig_28255_np12.3	41.34	5.62	12.64	40.41
*AdBGLU*04	>evm.model.contig_52431_np12.2	42.47	4.63	19.31	33.59
*AdBGLU*05	>evm.model.contig_52794_np12.2	39.06	6.05	16.99	37.89
*AdBGLU*06	>evm.model.contig_4149_np12.12	36.99	6.5	16.87	39.63
*AdBGLU*07	>evm.model.contig_3487_np12.4	40.09	4.84	12.95	42.12
*AdBGLU*08	>evm.model.contig_2826_np12.18	35.70	7.72	18.16	38.41
*AdBGLU*09	>evm.model.contig_6813_np12.9	39.36	6.97	16.20	37.48
*AdBGLU*10	>evm.model.contig_15195_np12.3	53.7	9.26	14.81	22.22
*AdBGLU*11	>evm.model.contig_5102_np12.18	41.38	6.90	24.14	27.59
*AdBGLU*12	>evm.model.contig_41190_np12.1	37.26	7.58	17.26	37.89
*AdBGLU*13	>evm.model.contig_51569_np12.1	41.77	6.83	13.65	37.75
*AdBGLU*14	>evm.model.scaffold_848_np12.18	39.92	5.14	18.58	36.36
*AdBGLU*15	>evm.model.contig_13620_np12.1	46.56	6.35	13.76	33.33
*AdBGLU*16	>evm.model.contig_875_np12.20	39.35	6.45	16.13	38.06
*AdBGLU*17	>evm.model.contig_5591_np12.41	45.21	4.11	6.85	43.84
*AdBGLU*18	>evm.model.contig_5591_np12.42	42.5	11.25	18.75	27.5
*AdBGLU*19	>evm.model.contig_7151_np12.13	38.85	8.92	17.2	35.03
*AdBGLU*20	>evm.model.contig_4524_np12.5	54.84	3.23	12.9	29.03
*AdBGLU*21	>evm.model.contig_6554_np12.21	51.92	15.38	11.54	21.15
*AdBGLU*22	>evm.model.contig_8631_np12.4	36.15	6.55	18.18	39.11
*AdBGLU*23	>evm.model.contig_8631_np12.5	38.42	7.53	15.06	39.00
*AdBGLU*24	>evm.model.contig_26865_np12.1	22.73	10.61	15.91	50.76
*AdBGLU*25	>evm.model.contig_8414_np12.6	35.51	7.48	22.43	34.58
*AdBGLU*26	>evm.model.contig_9063_np12.8	28.83	8.9	25.27	37.01
*AdBGLU*27	>evm.model.contig_16310_np12.17	41.07	4.55	13.64	40.75
*AdBGLU*28	>evm.model.contig_5761_np12.10	40.12	5.87	17.22	36.79
*AdBGLU*29	>evm.model.contig_4290_np12.7	17.57	12.16	25.68	44.59
*AdBGLU*30	>evm.model.contig_9965_np12.1	45.10	9.80	15.69	29.41
*AdBGLU*31	>evm.model.contig_5955_np12.1	34.09	7.32	15.91	42.68
*AdBGLU*32	>evm.model.contig_62681_np12.3	36.24	8.28	18.71	36.77
*AdBGLU*33	>evm.model.contig_3403_np12.17	17.36	9.92	34.71	38.02
*AdBGLU*34	>evm.model.contig_15768_np12.11	37.67	8.09	17.16	37.08
*AdBGLU*35	>evm.model.contig_9908_np12.3	42.86	5.19	24.68	27.27
*AdBGLU*36	>evm.model.contig_20919_np12.2	43.39	5.76	18.98	31.86
*AdBGLU*37	>evm.model.contig_20919_np12.4	41.43	5.61	20.25	32.71
*AdBGLU*38	>evm.model.contig_3686_np12.4	37.65	6.67	17.06	38.63
*AdBGLU*39	>evm.model.contig_4890_np12.4	41.91	7.59	11.55	38.94
*AdBGLU*40	>evm.model.contig_10929_np12.6	49.56	4.82	21.05	24.56
*AdBGLU*41	>evm.model.contig_40130_np12.2	21.18	5.88	23.53	49.41
*AdBGLU*42	>evm.model.contig_9053_np12.1	38.42	3.95	11.3	46.33
*AdBGLU*43	>evm.model.contig_4791_np12.1	41.03	7.69	21.79	29.49
*AdBGLU*44	>evm.model.contig_5084_np12.19	38.55	7.44	17.81	36.20
*AdBGLU*45	>evm.model.scaffold_6091_np12.6	40.52	2.60	11.15	45.72

保守域分析结果表明（图 6）Motif 8 为最短，含有 29 个氨基酸残基；Motif 6 稍长，含 35 个氨基酸残基；Motif 2、Motif 3 和 Motif 7 较长，含有 41 个氨基酸残基；Motif 1、Motif 4、Motif 5 最长，均含有 50 个氨基酸残基。通过保守基序结构可看出 Motif 5 的保守性较高。通过保守域分析发现，不同基因含有的保守域数量不同，在所有基序中，Motif 1 出现的频率最高，推测其为特征基序。



Motif 图中字母越高，表明该氨基酸出现的频率越大，则序列较为保守。

The higher the letter in the Motif diagram, the more frequently the amino acid appears, and the more conserved the sequence.

图 6 AdBGLU 家族蛋白的保守基序分析

Fig.6 Conserved motif analysis of AdBGLU family proteins

基于杭白芷和拟南芥的蛋白序列构建系统发育树（图 7），*AdBGLU* 基因被分为 6 个亚家族（A~F），*AdBGLU* 和 *AtBGLU* 基因同时存在于 B~F 亚族中，表明这些亚族中基因功能保守(张曼等，2022)。A 亚族中，有 3 个 *AtBGLU*，无 *AdBGLU*；B 亚族有 1 个 *AdBGLU* 和 4 个 *AtBGLU*；C 亚族有 13 个 *AdBGLU* 和 14 个 *AtBGLU*；D 亚族有 5 个 *AdBGLU* 和 8 个 *AtBGLU*；E 亚族有 14 个 *AdBGLU* 和 17 个 *AtBGLU*；F 亚族有 12 个 *AdBGLU* 和 2 个 *AtBGLU*。在 C 亚族中，杭白芷和拟南芥的基因数量相似，推测此亚族中的同源基因在拟南芥和杭白芷中可能发挥相似的作用(刘雨轩，2020)；而在其余亚族中，数量差异较大，可能存在调控杭白芷内香豆素合成的关键基因，此结论还需进一步验证。

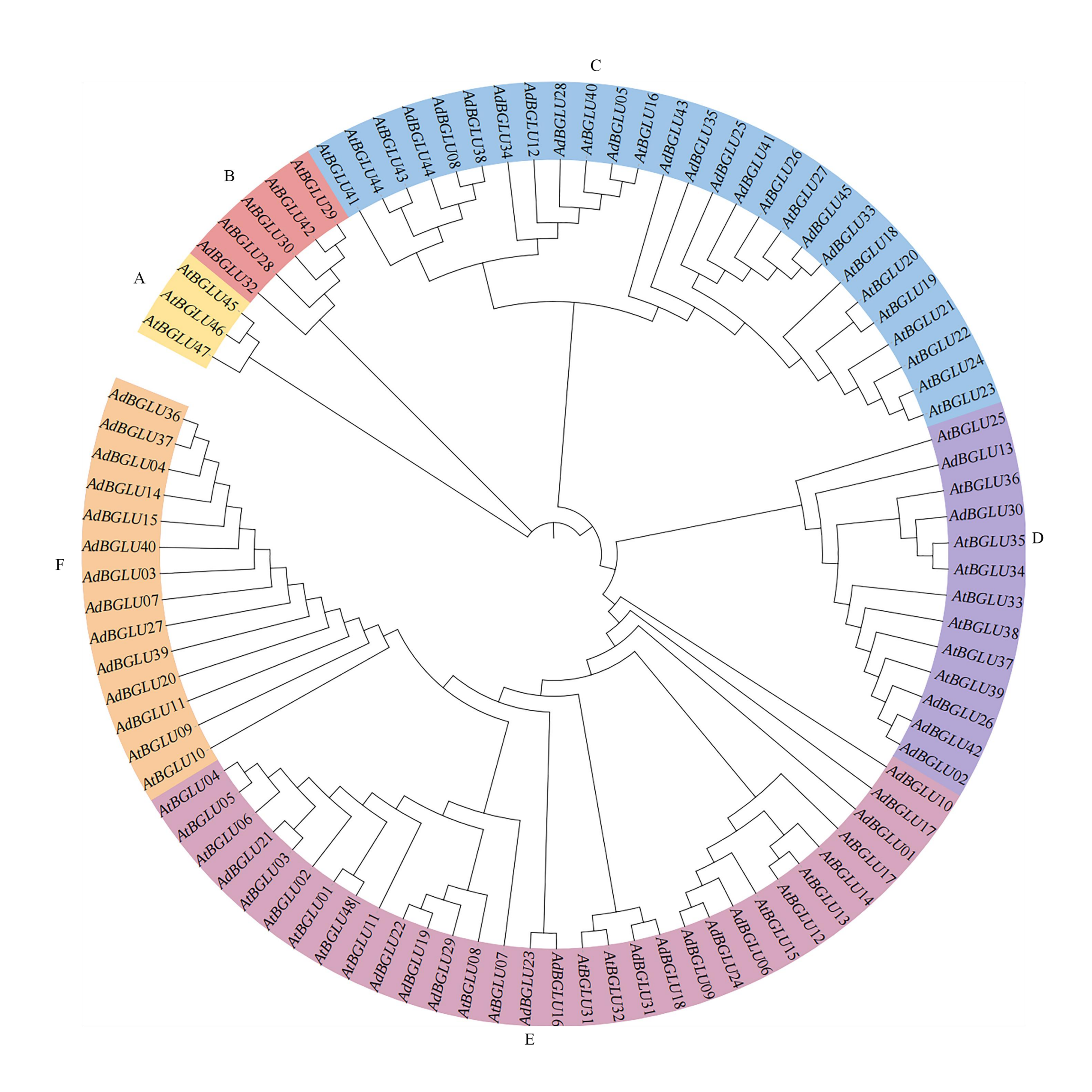


图 7 杭白芷和拟南芥 *BGLU* 基因家族进化分析

Fig.7 Phylogenetic analysis of *BGLU* proteins in *Angelica dahurica* and *Arabidopsis thaliana*

3 讨论与结论

有研究表明物种的基因组大小与其倍性水平及相应的染色体数目存在一定的正相关性(Mank & Avise, 2006)，通过对禾本科 282 种植物基因参数的研究发现，随着染色体倍性从二倍体到八倍体之间增加，其对应的基因组大小也显著增大，其基因组大小与倍性、染色体数呈极显著正相关(李桂双等，2012)。本研究获得约为 5.6 Gb 的杭白芷基因组，其他已完成基因组测序的伞形科植物有积雪草（约为 430 Mb）、芹菜（约为 3.33 Gb）、当归（约为 2.37Gb）(Han et al., 2022)、水芹（约为 1.28 Gb）、北柴胡（约为 621.42 Mb）、胡萝卜（约为 421.5 Mb）、野胡萝卜（约为 371.6 Mb）、芫荽（约为 2 130.29 Mb），其中，白芷、芹菜、当归、芫荽的染色体数目为 2n=22 条，积雪草和胡萝卜、野胡萝卜的染色体数目为 2n=18 条，北柴胡的染色体数目为 2n=12 条，除北柴胡外，符合染色体数目与基因组大小呈正相关关系，表明本次测得的杭白芷基因组大小符合染色体数目。白芷、芹菜的植株生长可达 1.5 m，而其余植物均不超过 1 m，初步推测伞形科植物基因组大小与植株高度呈正相关关系(邵晨等，2021)，可为后续同属或同科植物基因组的研究提供参考。

香豆素类化合物是一类具有重要药用价值的天然化合物，分为简单香豆素、呋喃香豆素、吡喃香豆素和其他香豆素四类 (王荣香等，2022)在植物中，香豆素通过苯丙烷代谢途径进行合成，目前已有较多研究揭示参与该生物合成途径的关键基因。例如，从明亮发光杆菌中提取的 *PAL* 基因能将 L-苯丙氨酸转化为肉桂酸、将 L-酪氨酸转化为对香豆酸 (ZHANG et al., 2021)；在对向日葵的研究中发现，有 3 个 *C4H* 基因具有催化肉桂酸生成对香豆酸，用同样方法对白花前胡和紫花前胡的 *C4H* 基因功能进行探索，发现都具有相同的催化功能(WANG et al., 2020)；在白花草木樨的研究中也发现 *MaBGLU1* 基因对于东莨菪苷形成东莨菪内酯具有关键作用(WU et al., 2022)；在白芷同属植物当归的研究中，发现 *PT* 基因对于呋喃香豆素的形成可能起到关键决定作用。*PAL*、*C4H* 等在香豆素生物合成途径中属于较为上游的基因，对于此类基因的研究较多，但是相对下游的 *BGLU* 基因的研究较少，尤其在白芷中更为缺乏。研究表明 *BGLU* 通过激活植物激素和防御化合物，与植物生理过程中的多个方面有关，尤其是对生物和非生物胁迫的响应。如陆地棉中 5 个 *GhBGLU* 或能正向调控棉花黄萎病抗性，拟南芥中的 *AtBGLU*10 可以催化游离 ABA 的产生，*AtBGLU*21-23 调控根中东莨菪苷的水解，*AtBGLU*42 参与诱导机体对细胞疾病的抵抗力。本研究所获得的杭白芷基因组，能为后续进行白芷中香豆素类成分合成相关基因的挖掘提供基础，具有重要价值及意义。

目前，已在拟南芥中发现 48 个 *BGLU* 家族基因，玉米中发现 26 个(Gómez-Anduro et al., 2011)，水稻中发现 40 个(Opassiri et al., 2006)，大豆中发现 42 个(柯丹霞等，2019)，陆地棉中发现 53 个(张曼等，2022)，苜蓿发现 51 个(Yang et al., 2021)，本研究在杭白芷中鉴定出 45 个 *BGLU* 家族基因，并对其进行理化性质、二级结构等分析，发现其亚细胞定位多在细胞质、叶绿体、液泡中，这一结论与玉米中的β-葡萄糖苷酶定位基本一致(Kristoffersen et al., 2000)，*AdBGLU* 基因家族的理化性质、二级结构、亚细胞定位等特征差异较大，说明该基因家族的结构较为复杂，推测其功能较为多样，各基因在功能分工上有所不同，在生物体内参与多种不同代谢过程。杭白芷中存在多种香豆素类化合物，如欧前胡素、异欧前胡素、白当归素、佛手柑内酯等等，其生物合成途径也较为复杂，这可能是与 *AdBGLU* 基因功能的多样有关。*AdBGLU* 的初步分析对杭白芷香豆素生物合成具有重要作用，可为进一步揭示和利用杭白芷香豆素类成分合成途径关键基因的功能提供前期基础。

4 数据获得

原始测序数据已上传至国家基因库生命大数据平台（CNGBdb，https://db.cngb.org/），项目编号为 CNP0003549。

参考文献：

DUAN Z, WU F, YAN Q, et al., 2022. Research progress on plant coumarin biosynthesis pathway and the genes encoding the key enzymes[J]. Acta Pratacult Sin, 31(1): 217-228.[段珍，吴凡，闫启，等，2022. 植物香豆素生物合成途径及关键酶基因研究进展[J]. 草业学报，31(1): 217-228.]

EDGAR RC, 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput[J]. Nucl Acids Res, 32(5): 1792-1797.

FINN RD, MISTRY J, SCHUSTER-BÖCKLER B, et al., 2006. Pfam: clans, web tools and services[J]. Nucl Acid Res, 34(Database issue): D247-D251.

GÓMEZ-ANDURO G, CENICEROS-OJEDA EA, CASADOS-VÁZQUEZ LE, et al., 2011. Genome-wide analysis of the beta-glucosidase gene family in maize (*Zea mays* L. var B73)[J]. Plant Mol Biol, 77(1-2): 159-183.

GUINDON S, GASCUEL O, 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood[J]. Syst Biol, 52(5): 696-704.

HAN X, LI C, SUN S, et al., 2022. The chromosome-level genome of female ginseng (*Angelica sinensis*) provides insights into molecular mechanisms and evolution of coumarin biosynthesis[J]. Plant J, 112(5): 1224-1237.

HUANG WJ, XU X, CHEN JS, et al., 2021. Bioinformatics analysis and expression pattern of NAC transcription factor family of *Angelica dahurica* var. *formosana* from Sichuan province[J]. Chin J Chin Mat Med, 46(7): 1769-1782. [黄文娟，许鑫，陈靳松，等，2021. 川白芷 NAC 家族的生物信息及表达模式分析[J]. 中国中药杂志，46(7)：1769-1782.]

IORIZZO M, ELLISON S, SENALIK D, et al., 2016. A high-quality carrot genome assembly provides new insights into carotenoid accumulation and asterid genome evolution[J]. Nat Genet, 48(6): 657-666.

JI Q, MA YH, ZHANG Y, 2020. Research progress on chemical constituents and pharmacological effects of *Angelicae dahuricae* radix[J]. Food Drug, 22(6): 509-514. [吉庆，马宇衡，张烨，2020. 白芷的化学成分及药理作用研究进展[J]. 食品与药品，22(6)：509-514.]

JIANG YJ, JIANG YM, YAO F, et al., 2021. Bioinformatics analysis on the CONSTANS-like protein family in *Angelica dahurica* var. *formosana*[J]. Mol Plant Breed, 19(12): 3923-3931. [蒋翼杰，江美彦，姚菲，等，2021. 川白芷 CONSTANS-like 蛋白家族生物信息学分析[J]. 分子植物育种，19(12)：3923-3931.]

KE DX, LIU YH, ZHANG JJ, et al., 2019. Genome-wide identification and expression analysis of BGLU family genes in Soybean[J]. J Xinyang Norm Univ(Nat Sci Ed), 32(3): 372-378. [柯丹霞，刘永辉，张静静，等，2019. 大豆 BGLU 基因家族全基因组鉴定与表达分析[J]. 信阳师范学院学报(自然科学版)，32(3)：372-378.]

KOREN S, WALENZ BP, BERLIN K, et al., 2017. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation[J]. Genome Res, 27(5): 722-736.

KRISTOFFERSEN P, BRZOBOHATY B, HÖHFELD I, et al., 2000. Developmental regulation of the maize Zm-p60.1 gene encoding a beta-glucosidase located to plastids[J]. Planta, 210(3): 407-415.

LARKIN MA, BLACKSHIELDS G, BROWN NP, et al., 2007. Clustal W and Clustal X version 2.0[J]. Bioinformatics, 23(21): 2947-2948.

LI B, ZHANG X, WANG J, et al., 2014. Simultaneous characterisation of fifty coumarins from the roots of *Angelica dahurica* by off-line two-dimensional high-performance liquid chromatography coupled with electrospray ionisation tandem mass spectrometry[J]. Phytochem Analysis, 25(3): 229-240.

LI GS, CAO B, BAI CK, 2012. Correlation analysis between genome size and seed characteristics in poaceae plants[J]. Bull Bot Res, 32(6): 701-706. [李桂双，曹博，白成科，2012. 禾本科植物基因组大小与种子特性的相关性分析[J]. 植物研究，32(6)：701-706.]

LI L, STOECKERT CJ Jr, ROOS DS, 2003. OrthoMCL: identification of ortholog groups for eukaryotic genomes[J]. Genome Res, 13(9): 2178-2189.

LIU YX, 2020. Identification and expression analysis of WRKY gene family in *Solanum lycopersicum*[D]. Shenyang: Shenyang Agricultural University: 1-79. [刘雨轩，2020. 番茄 WRKY 基因家族成员鉴定及表达分析[D]. 沈阳：沈阳农业大学：1-79.]

LIU Y, 2019. Studies on bacteriostatic mechanism of *Angelica dahurica* and excavationof key genes of coumarin biosynthesis[D]. Chengdu: Sichuan Agricultural University: 1-69. [刘洋，

2019. 川白芷抑菌机理研究及香豆素生物合成关键基因的挖掘[D]. 成都：四川农业大学：1-69.]

MANK JE, AVISE JC, 2006. Cladogenetic correlates of genomic expansions in the recent evolution of actinopterygiian fishes[J]. Proceed Royal Soc B Biol Sci, 273(1582):33-38.

NATIONAL PHARMACOPOEIA COMMISSION, 2020. Pharmacopoeia of People's Republic of China: 1[M]. Beijing: China Medical Science Press: 109-110. [国家药典委员会，2020. 中华人民共和国药典：一部[M]. 北京：中国医药科技出版社：109-110.]

OPASSIRI R, POMTHONG B, ONKOKSOONG T, et al., 2006. Analysis of rice glycosyl hydrolase family 1 and expression of Os4bglu12 beta-glucosidase[J]. BMC Plant Biol, 6: 33.

SAMPEDRO J, VALDIVIA ER, FRAGA P, et al., 2017. Soluble and membrane-bound β-glucosidases are involved in trimming the xyloglucan backbone[J]. Plant Physiol, 173(2): 1017-1030.

SENOL CALI D, KIM JS, GHOSE S, et al., 2019. Nanopore sequencing technology and tools for genome assembly: computational analysis of the current state, bottlenecks and future directions[J]. Brief Bioinform, 20(4): 1542-1559.

SHA LP, 2018. Examples of CTAB method, SDS method and salting-out method for crude extraction of plant DNA[J]. Teach Middle Sch Biol, 21: 65-67. [沙丽萍，2018. 例谈植物DNA粗提取的CTAB法、SDS法与盐析法[J]. 中学生物教学，21：65-67.]

SHAO C, LI YQ, LUO A, et al., 2021. Relationship between functional traits and genome size variation of angiosperms with different life forms[J]. Biodivers Sci, 29(5): 575-585. [邵晨，李耀琪，罗奥，等，2021. 不同生活型被子植物功能性状与基因组大小的关系[J]. 生物多样性，29(5)：575-585.]

SIMÃO FA, WATERHOUSE RM, IOANNIDIS P, et al., 2015. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs[J]. Bioinformatics, 31(19): 3210-3212.

SONG X, WANG J, LI N, et al., 2020. Deciphering the high-quality genome sequence of coriander that causes controversial feelings[J]. Plant Biotechnol J, 18(6): 1444-1456.

SONG X, SUN P, YUAN J, et al., 2021. The celery genome sequence reveals sequential paleo-polyploidizations, karyotype evolution and resistance gene reduction in apiales[J]. Plant Biotechnol J, 19(4): 731-744.

SUN HH, XUE YM, LIN YF, 2014. Enhanced catalytic efficiency in quercetin-4'-glucoside hydrolysis of *Thermotoga maritima* β-glucosidase A by site-directed mutagenesis[J]. J Agric Food Chem, 62(28): 6763-6770.

VENUGOPALA KN, RASHMI V, ODHAV B, 2013. Review on natural coumarin lead compounds for their pharmacological activity[J]. Biomed Res Int, 2013: 963248.

WANG R, LIU J, YANG DY, et al., 2020. Research progress in chemical constituents and pharmacological action of *Angelica dahurica*[J]. Inf Trad Chin Med, 37(2): 123-128. [王蕊，刘军，杨大宇，等，2020. 白芷化学成分与药理作用研究进展[J]. 中医药信息，37(2)：123-128.]

WANG RX, SONG J, SUN B, et al,. 2022. Research progress of function and biosynthesis of coumarins[J]. Chin Biotechnol, 42(12): 79-90. [王荣香，宋佳，孙博，等，2022. 香豆素类化合物功能及生物合成研究进展[J]. 中国生物工程杂志，42(12)：79-90.]

WANG Z, JIAN X, ZHAO Y, et al.,2020. Functional characterization of cinnamate 4-hydroxylase from *Helianthus annuus* Linn using a fusion protein method[J]. Gene, 758: 144950.

WILKINS MR, GASTEIGER E, BAIROCH A, et al., 1999. Protein identification and analysis tools in the ExPASy server[J]. Meth Mol B, 112: 531-552.

WU F, DUAN Z, XU P, et al., 2022. Genome and systems biology of *Melilotus albus* provides insights into coumarins biosynthesis[J]. Plant Biotechnol J, 20(3): 592-609.

WU F, 2021. Study on whole genome sequencing and functional genes of key traits in *Cleistogenes songorica* and *Melilotus albus*[D]. Lanzhou: Lanzhou University: 1-185. [吴凡，2021. 无芒隐子草和白花草木樨全基因组及其关键性状相关功能基因研究[D]. 兰州：兰州大学：1-185.]

WU P, GUO JX, WANG XY, et al., 2020. High-throughput transcriptome sequencing of roots of *Angelica dahurica* and data analyses[J]. Mol Plant Breed, 2020, 18(10): 3207-3216.[吴萍，郭俊霞，王晓宇，等，2020. 基于高通量测序技术的杭白芷(*Angelica dahurica*)根转录组数据分析[J]. 分子植物育种，18(10)：3207-3216.]

XU Z, WANG H, 2007. LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons[J]. Nucl Acid Res, 35(Web Server issue): W265-W268.

YANG J, MA L, JIANG W, et al., 2021. Comprehensive identification and characterization of abiotic stress and hormone responsive glycosyl hydrolase family 1 genes in *Medicago truncatula*[J]. Plant Physiol Biochem, 158: 21-33.

YAO F, JIANG MY, YANG YS, et al., 2022. Bioinformatics and expression analysis on MYB-related family in *Angelicae dahuricae* var. *formosana*[J]. Chin J Chin Mat Med, 47(7): 1831-1846. [姚菲，江美彦，杨云舒，等，2022. 川白芷 MYB-related 家族的生物信息及表达模式分析[J]. 中国中药杂志，47(7)：1831-1846.]

YU KP, PENG C, LIN YL, et al., 2023. Expression of β-glucosidase An-bgl3 from *Aspergillus niger* for conversion of scopoline[J]. Chin J Biotechnol, 39(3): 1232-1246. [于坤朋，彭程, 林燕玲等, 2023. 黑曲霉β-葡萄糖苷酶 An-bgl3 的重组表达及东莨菪苷的转化[J]. 生物工程学报，39(3)：1232-1246.]

YU J, ZHU YH, 2014. Summary of the application of *Angelica dahurica* in ancient prescription[J]. Heilongjiang Med J, 27(1): 156-158. [于静，朱艳华，2014. 中药白芷在古方中美白作用的应用概述[J]. 黑龙江医药，27(1)：156-158.]

ZHANG F, REN J, ZHAN J, 2021. Identification and characterization of an efficient phenylalanine ammonia-lyase from *Photorhabdus luminescens*[J]. Appl Biochem Biotechnol, 193(4): 1099-1115.

ZHANG M, WANG ZC, LIU ZW, et al., 2022-09-17. Genome-wide identification and analysis of BGLU genes family in *Gossypium hirsutum*[J/OL]. J Agric Sci Technol: 1-12. [张曼，王志城，刘正文，等，2022-09-17. 陆地棉 BGLU 基因家族成员的全基因组鉴定与表达分析[J/OL]. 中国农业科技导报：1-12.]

ZHAO H, FENG YL, WANG M, et al., 2022. The *Angelica dahurica*: a review of traditional uses, phytochemistry and pharmacology[J]. Front Pharmacol, 13: 896637.